Data Mining
ID2222

Homework 1: Finding Similar Items: Textually Similar Documents

Authors:
Abdul Aziz Alkathiri
Francesco Staccone
-

November 17, 2019

1 Introduction

For this assignment, we have chosen to use the Enron Email dataset (downloaded from: https://www.kaggle.com/wcukierski/enron-email-dataset), which contains 500,000+ instances of emails generated by employees of the Enron Corporation.

The dataset contains two columns, the first of which is named "file" which is a unique identifier for each email, and the second is "message", which is the content of the emails themselves, as shown in Figure 1. We are mostly interested in the "messages" part of the dataset so we dropped the first column and filtered the content of the emails such that it only includes the relevant email content, and kept only 500 instances (see Figure 2).

	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

Figure 1: Full dataset.

	1
0	Here is our forecast
1	Traveling to have a business meeting takes the fun out of the trip. Especially if you have to prepare a presentation. I would suggest holding the business plan meetings here then take a trip without any formal business meetings. I would even try and get some honest opinions on whether a trip is even desired or necessary.As far as the business meetings, I think it would be more productive to try and stimulate discussions across the different groups about what is working and what is not. Too often the presenter speaks and the others are quiet just waiting for their turn. The meetings might be better if held in a round table discussion format. My suggestion for where to go is Austin. Play golf and rent a ski boat and jet ski's. Flying somewhere takes too much time.
2	test successful. way to go!!!
3	Randy, Can you send me a schedule of the salary and level of everyone in the scheduling group. Plus your thoughts on any changes that need to be made. (Patti S for example)Phillip
4	Let's shoot for Tuesday at 11:45.

Figure 2: Cleaned dataset of the first 500 instances with relevant contents retained.

Our submission consists of two .ipynb files, one for the data cleaning and the other for running the tasks of the assignment. Included also are the cleaned dataset files "dropped.csv" and "filtered.csv": the first one contains the first 500 emails in the dataset, while the second one contains the first 500 emails among the longest ones. The dataset containing the *longest* 500 emails is used for our anylisis. The original dataset is not included due to its large size but it is referred to above. We feel like submitting notebooks is more easier and more intuitive to replicate the steps we took.

2 Implementation

Both in pre-processing and tasks implementation, we used pandas library. We first implemented shingling (picking shingles of size $k = 5$). To compress shingle sizes, they are hashed to fit in integers, as shown in Figure 3. Then shingles from each email (observation) are appended to a universal list which contains all of the shingles. For the second task, we also measured the Jaccard similarities of the emails as similarities among sets of shingles, as can be seen in Figure 4. For instance, if one were to look at the content of two emails with Jaccard similarity of 1, say, emails 0 and 22, they would find the contents identical.

```
[213316875, 284929611, 638525431, 676430235, 1337481804, 1609821938, 2007683970, 2062840440, 2752276938, 3065203866, 31
06968482, 3181387989, 3519895912, 3603488944, 3770746557, 3883998174, 4113972070]
[60831, 248770, 5011777, 6547551, 33848650, 35799293, 49326850, 49659464, 49874867, 54997773, 61033506, 67665585, 81262
134, 84427501, 86979466, 94512332, 109515330, 109915699, 112797060, 115974342, 120340269, 135119813, 135554313, 1357864
50, 137426469, 137681209, 142818482, 145191191, 148405964, 148690772, 156393036, 163192231, 181089256, 184246633, 19346
1049, 194083427, 196463189, 202003542, 207306152, 207481743, 208730773, 220243434, 243690175, 246655464, 250370747, 255
505537, 259676759, 264106082, 264627047, 270667761, 290346326, 293235070, 299535211, 305718398, 305753316, 307606693, 3
07979160, 364907921, 371202685, 371743592, 386518964, 392828419, 402302711, 411507155, 413823119, 436243395, 437580203,
442149599, 459699912, 462842152, 464590535, 467599722, 469990038, 473693122, 480614478, 495948614, 498626934, 50016812
1, 501470986, 508227356, 518546981, 537857638, 550928451, 552609583, 553680600, 559579786, 578137199, 579227695, 579339
434, 580554734, 589839017, 594963263, 604525109, 608231713, 618613449, 621531882, 625984329, 628415777, 637951864, 6467
58172, 647150407, 648832350, 658146159, 659945128, 666542161, 667755981, 687136356, 688086009, 725039679, 730707434, 73
2143381, 738290926, 745604822, 755001209, 757834738, 765932119, 773986885, 778226605, 779989422, 780974751, 786562423,
791078491, 792278640, 806876995, 814833533, 833450248, 833866022, 844484067, 846675806, 847153406, 848543145, 84861859
5, 848835131, 851638015, 861367142, 862126322, 863212951, 879581510, 880779359, 884726699, 888836052, 890774302, 895764
605, 901026131, 904104312, 906854221, 912958646, 913427973, 927976984, 939465437, 942135126, 948972786, 962403650, 9631
83025, 963447298, 981031714, 982192673, 983333598, 996037398, 997791337, 1007378858, 1009336484, 1022202606, 102978885
3, 1031090762, 1047124071, 1060636614, 1062438805, 1063858464, 1080657545, 1080730940, 1091194744, 1102749324, 11131298
73, 1114357072, 1130744506, 1131636448, 1138902765, 1141260155, 1145543184, 1174107612, 1177403790, 1180087470, 1184958
```

Figure 3: Hashed shingles, where each list represents the shingles of each email.

```
 + SHINGLES +

(mail 0, mail 1): 0.9963963963963964
(mail 0, mail 22): 1.0
(mail 0, mail 23): 0.9963963963963964
(mail 0, mail 67): 0.9963963963963964
(mail 0, mail 68): 1.0
(mail 0, mail 115): 1.0
(mail 0, mail 116): 0.9963963963963964
(mail 1, mail 22): 0.9963963963963964
(mail 1, mail 23): 1.0
(mail 1, mail 67): 1.0
(mail 1, mail 68): 0.9963963963963964
(mail 1, mail 115): 0.9963963963963964
(mail 1, mail 116): 1.0
(mail 2, mail 3): 0.9980963259090044
(mail 2, mail 19): 0.9956298688960669
(mail 2, mail 24): 1.0
(mail 2, mail 25): 0.9980963259090044
```

Figure 4: Computing the Jaccard similarity of two sets if hashed shingles - only sets that are above a threshold (0.5) are printed.

Next down the line we implemented minHashing but instead of using permutations, we implemented minHashing using hash functions with the number of hashes k=100, the resulting Signature Matrix is shown in Figure 5.

```
In [107]: print (signature_matrix)
          print (signature_matrix.shape[1])
          print (signature_matrix.shape[0])

[[ 236984  236984  550829 ...  236984  739005  128082]
 [6461968 6461968 1174424 ...  174246  595905  595905]
 [1127943 1127943  334398 ...  603424 1512633  157603]
 ...
 [ 506232  506232  126608 ...  166434  537334  299129]
 [2834679 2834679 1589689 ...   89157 2742058 2742058]
 [1179936 1179936  937744 ...  839607  700387  279796]]
500
100
```

Figure 5: Signature matrix, with its dimensions displayed.

We measured the similarity of two integer vectors of the Signature Matrix as a fraction of components, in which they agree, as shown in Figure 6. Likewise here, the contents of two emails with the similarity of 1 are identical.

Lastly, we implemented LSH where we produced candidate sets based on the parameters, i.e. the similarity threshold t, the number of bands b, and the number of rows r in each band.

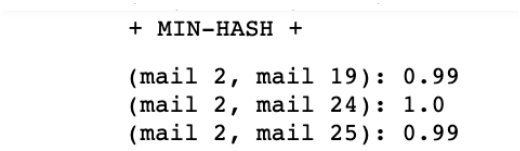

Figure 6: MinHash similarity, only similarities above a threshold (0.5) are printed.

These parameters can be changed. For instance, with the parameters of t= 0.9, b= 20 and r= 5, we get 729 candidate pairs, as shown in Figure 7. Of course, these parameters can be changed but one caveat must be observed: $b*r = k$.

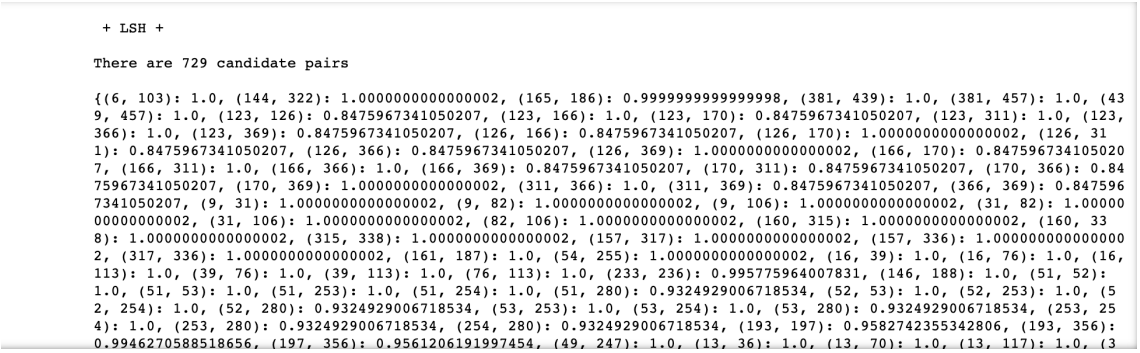

Figure 7: Candidate pairs, only pairs with similarity above a threshold (0.5) are printed.

The runtime of the whole process from Shingling to LHS for 500 emails takes 261.515 seconds.